KU LEUVEN

FACULTY OF ECONOMICS AND BUSINESS

Robust and sparse estimation of the inverse covariance matrix using rank correlation measures

Croux C, Öllerer V.

KBI_1512

Robust and sparse estimation of the inverse covariance matrix using rank correlation measures

Christophe Croux, Viktoria Öllerer

Faculty of Economics and Business, KU Leuven, Belgium

Spearman's rank correlation is a robust alternative for the standard correlation coefficient. By using ranks instead of the actual values of the observations, the impact of outliers remains limited. In this paper, we study an estimator based on this rank correlation measure for estimating covariance matrices and their inverses. The resulting estimator is robust and consistent at the normal distribution. By applying the graphical lasso, the inverse covariance matrix estimator is positive definite if more variables than observations are available in the data set. Moreover, it will contain many zeros, and is therefore said to be sparse. Instead of Spearman's rank correlation, one can use the Quadrant correlation or Gaussian rank scores. A simulation study compares the different estimators. This type of estimator is particularly usefull for estimating (inverse) covariance matrices in high dimensions, when the data may contain several outliers in many cells of the data matrix. More traditional robust estimators are not well defined or computable in this setting. An important feature of the proposed estimators is their simplicity and easyness to compute using existing software.

1 Introduction

We have a sample of n multivarate observations, and for each of these observations we measure p variables. The resulting data can be collected in a data matrix $\mathbf{X}$ where the observations are the rows of the data matrix, and each variable corresponds to a column of the data matrix. The data matrix $\mathbf{X}$ has np cells, where a cell contains a univariate measurement x_{ij}:

$$\mathbf{X} = \begin{pmatrix} x_{11} & \dots & \dots & x_{1p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & \dots & \dots & x_{np} \end{pmatrix}.$$

Typically, these data matrices are thin, with n much larger than p. But in this paper focus is on fat data matrices with more columns than rows. Fat data

matrices often occur in practice. For instance in medicine where hundreds of variables are measured for a limited set of patients. The transposed rows of $\mathbf{X}$ are denoted as $\mathbf{x}_1, \ldots, \mathbf{x}_n$, with $\mathbf{x}_i \in \mathbb{R}^p$. The columns of the data matrix are denoted as $\mathbf{x}^1, \ldots, \mathbf{x}^p \in \mathbb{R}^n$.

We assume that the observations are a random sample of a multivariate normal distribution with mean μ and covariance matrix $\mathbf{\Sigma}$. This covariance matrix is assumed to be positive definite, hence, all its eigenvalues are strictly positive. The aim is to estimate the unknown parameters μ and $\mathbf{\Sigma}$ from the data such that (i) the estimators are resistant to outlying cells (ii) the estimate of $\mathbf{\Sigma}$ is positive definite.

In high dimensions, the occurrence of outliers is to be expected. Data are collected less carefully, often in an automatic and inaccurate way. Gross-errors can occur. Moreover, the size of the data set and the large number of variables makes outlier detection using visualization cumbersome. Therefore, estimators should be robust to outlying values x_{ij}, hence outlying cells. In the traditional literature on robust statistics [see Maronna et al., 2006, for a more recent textbook], one considers outlying observations, and an observation is already an outlier if only one of its cells is outlying. In high dimensions, the notion of outlying cells is more appropriate. Indeed, take $p = 200 > n = 100$ and assume that every cell x_{ii}, for $1 \leq i \leq n$ is an outlier. Then all observations are outliers, suggesting that robust estimation would be impossible. But only 0.5% of the cells are outliers! In Section 4 the concept of breakdown point under cellwise contamination, as introduced in Öllerer and Croux [2014], is defined. The estimators advocated in this paper have a high breakdown point according to this definition, showing that robust estimators do exist in high dimensions. One only needs to reconsider what appropriate measures for robustness are in high dimensions.

The sample covariance matrix estimator

$$\hat{\mathbf{\Sigma}} = \frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top \tag{1}$$

with $\bar{\mathbf{x}}$ the sample average is not only non-robust, it also has the problem that it is only positive semidefinite. Some of its eigenvalues will be zero if $p \geq n$. Hence, its inverse is not existing. In multivariate statistics one often needs the inverse: to compute Mahalanobis distances, for Fisher discriminant analysis, ... Therefore, we want to have an estimator of $\mathbf{\Sigma}$ that is always positive definite. We can achieve this in many ways, but a popular choice is the Graphical Lasso, or Glasso of Friedman et al. [2008]. Glasso takes a positive semidefinite covariance matrix estimator as an input, and returns a positive definite one. A particular feature of Glasso is that the resulting estimator of the inverse covariance matrix is also sparse, meaning that many of its element are exactly equal to zero. We denote the inverse covariance matrix, or *precision matrix*, as $\mathbf{\Theta} = \mathbf{\Sigma}^{-1}$.

In Section 2 we define the estimators of the precision matrix to be studied. They are robust to cellwise outliers, and give sparse and positive definite estimates of $\mathbf{\Theta}$. In Section 3 we give some R code to show how easily the estimates can be computed. Theoretical results are presented in Section 4. The different estimators are compared in Section 5 by means of a simulation experiment. Section 6 shows how the estimators have been used for Graphical modelling. Section 7 contains some final discussion.

2 Estimators

We follow the approach of Tarr et al. [2015] for constructing sparse and robust precision matrices. In a first step, we construct a robust estimator $\mathbf{S}$ of the covariance matrix. In a last step, $\mathbf{S}$ serves as an input of Glasso, resulting in a sparse and robust estimator $\hat{\mathbf{\Theta}}_{\mathbf{S}}$ of the precision matrix:

$$\hat{\mathbf{\Theta}}_{\mathbf{S}} = \underset{\substack{\mathbf{\Theta}=(\theta_{jk})\in\mathbb{R}^{p\times p}\\ \mathbf{\Theta}\succ 0}}{\arg\max} \log\det(\mathbf{\Theta}) - \operatorname{tr}(\mathbf{S}\mathbf{\Theta}) - \lambda \sum_{j,k=1}^{p} |\theta_{jk}|, \tag{2}$$

where the maximization is over all positive definite matrices $\mathbf{\Theta} \succ 0$. The algorithm for solving (2) requires the input matrix $\mathbf{S}$ to be symmetric and positive semidefinite. A stable implementation of Glasso is given in the R-package `huge` [Zhao et al., 2014a]. The parameter λ in (2) controls for the sparsity of the solution: the larger λ, the sparser the precision matrix estimate. We compute $\hat{\mathbf{\Theta}}_{\mathbf{S}}$ over a logarithmic spaced grid of ten values, as is done by default in the `huge`-package. The final solution is then the one with lowest value of the following Bayesian Information Criterion [see Yuan and Lin, 2007]:

$$BIC(\lambda) = -\log\det\hat{\mathbf{\Theta}}_{\mathbf{S}} + \operatorname{tr}(\hat{\mathbf{\Theta}}_{\mathbf{S}}\mathbf{S}) + \frac{\log n}{n}\sum_{i\le j}\hat{e}_{ij}(\lambda). \tag{3}$$

with $\hat{e}_{ij} = 1$ if $(\hat{\mathbf{\Theta}}_{\mathbf{S}})_{ij} \neq 0$ and $\hat{e}_{ij} = 0$ otherwise. Note that $\hat{\mathbf{\Theta}}_{\mathbf{S}}$ depends on λ.

2.1 Two-step Estimators

So how do we choose $\mathbf{S}$? Tarr et al. [2015] propose to use the robust covariance of Gnanadesikan and Kettenring [1972] between $\mathbf{x}^j$ and $\mathbf{x}^k$ for s_{jk}, with s_{jk} an element of $\mathbf{S}$. Öllerer and Croux [2014] showed that this choice leads to some loss of robustness and a too high computational cost. Instead they propose to use

$$s_{jk} = \operatorname{scale}(\mathbf{x}^j)\operatorname{scale}(\mathbf{x}^k) r(\mathbf{x}^j, \mathbf{x}^k) \qquad j,k = 1,\ldots,p. \tag{4}$$

As scale estimator scale() the robust Q_n-estimator [Rousseeuw and Croux, 1993] is taken, which has the highest possible breakdown point of all scale estimator and is quite efficient at the normal model. For the correlation $r(\mathbf{x}^j, \mathbf{x}^k)$ Öllerer and Croux [2014] considered the following three choices.

- The Quadrant correlation, defined as

$$r_{Quadrant}(\mathbf{x}^j, \mathbf{x}^k) = \frac{1}{n}\sum_{i=1}^{n} \text{sign}((x_{ij} - \text{med}_{\ell=1,\ldots,n}\, x_{\ell j})(x_{ik} - \text{med}_{\ell=1,\ldots,n}\, x_{\ell k})), \tag{5}$$

  where sign$(\cdot)$ denotes the sign-function.

- The Spearman correlation defined as the sample correlation of the ranks of the observations:

$$r_{Spearman}(\mathbf{x}^j, \mathbf{x}^k) = \sum_{i=1}^{n} \frac{(R(x_{ij}) - \frac{n+1}{2})(R(x_{ik}) - \frac{n+1}{2})}{\sqrt{\sum_{i=1}^{n}(R(x_{ij}) - \frac{n+1}{2})^2 \sum_{i=1}^{n}(R(x_{ik}) - \frac{n+1}{2})^2}}, \tag{6}$$

  with $R(x_{ij})$ the rank of x_{ij} among all elements of $\mathbf{x}^j$, for any $1 \leq j \leq p$ and $1 \leq i \leq n$.

- The Gaussian rank correlation defined as the sample correlation estimated from the normal scores of the data:

$$r_{Gauss}(\mathbf{x}^j, \mathbf{x}^k) = \frac{\sum_{i=1}^{n} \Phi^{-1}(\frac{R(x_{ij})}{n+1})\Phi^{-1}(\frac{R(x_{ik})}{n+1})}{\sum_{i=1}^{n}(\Phi^{-1}(\frac{i}{n+1}))^2}, \tag{7}$$

  where $\Phi(\cdot)$ is the cumulative distribution function of a standard normal.

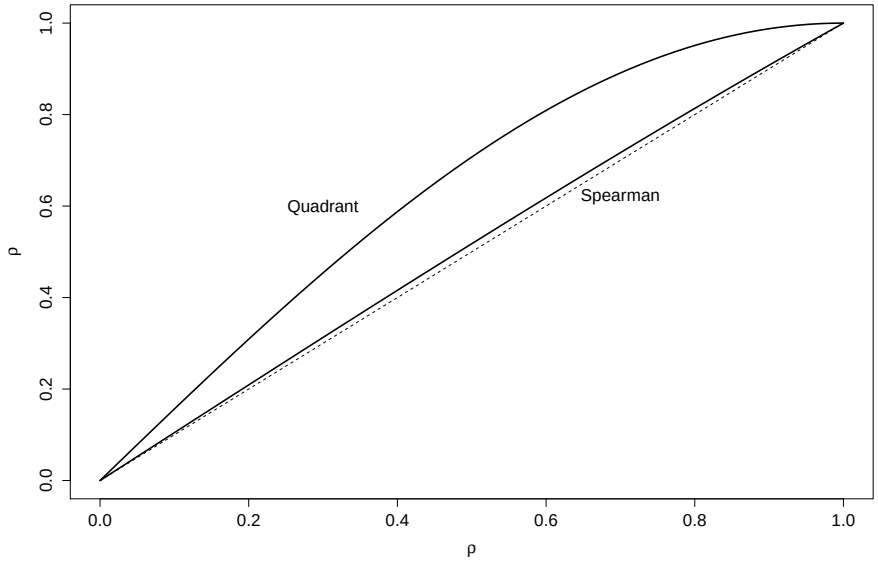
The robustness and efficiency properties of the Quadrant and Spearman correlation are studied in Croux and Dehon [2010], and of the Gaussian rank correlation in Boudt et al. [2012]. Using these correlation meausres, combined with (4), yields positive semidefinite covariance matrices.

2.2 Three-step Estimators

The Quadrant and Spearman correlation are not consistent at the bivariate normal distribution. This means that Quadrant and Spearman correlation between two variables having a joint normal distribution with correlation ρ do not estimate ρ, not even if the sample size is infinite. The corresponding $\mathbf{S}$ is not a consistent estimator of $\mathbf{\Sigma}$ and has an asymptotic bias. To resolve this inconsistency, the following transformations need to be applied:

$$\tilde{r}_{Quadrant} = \sin(\frac{\pi}{2} r_{Quadrant}) \tag{8}$$

Figure 1: Plot of the transformations $\rho \to \sin(\pi\rho/2)$ and $\rho \to 2\sin(\pi\rho/6)$ needed for making Spearman and Quadrant correlation consistent, together with the 45 degrees line.



and

$$\tilde{r}_{Spearman} = 2\sin(\frac{\pi}{6} r_{Spearman}). \tag{9}$$

Hence, to get consistency, the transformed Spearman and Quadrant correlation need to be plugged into (4). It is instructive to plot the transformations (8) and (9). We see from Figure 1 that the asymptotic bias of the Spearman correlation is very small; the transformation pushes the Spearman correlation only slightly upwards. On the other hand, the Quadrant correlation is more severely underestimating the population correlation ρ.

Unfortunately, the resulting $\mathbf{S}$ will not be positive semidefinite anymore, and cannot be used safely as input for Glasso. Therefore, an additional step to make $\mathbf{S}$ positive semidefinite is needed before Glasso can be applied. We implement two easy ways to do this, but other possibilities do exist [see Zhao et al., 2014b]. Denote λ_j and $\mathbf{v}_j$ the eigenvalues and eigenvectors of the matrix $\mathbf{S}$, respectively, for $1 \leq j \leq p$. Since $\mathbf{S}$ is symmetric, these eigenvalues exist as real numbers, but may be negative.

1. The perturbation method is an heuristical approach often used in regularization. One simply adds a non-negative value to all diagonal elements of $\mathbf{S}$:

$$\mathbf{S}_{perturb} = \mathbf{S} + |\min(0, \min_j \lambda_j)|\, \mathbf{I}. \tag{10}$$

It is immediate to see that the resulting covariance matrix has no negative eigenvalues any more.

2. Rousseeuw and Molenberghs [1993] proposed to use

$$\mathbf{S}_{npd} = \sum_{j=1}^{p} \max(0, \lambda_j) \mathbf{v}_j \mathbf{v}_j^t. \tag{11}$$

It has been show [e.g. Zhao, Roeder, and Liu, 2014b] that $\mathbf{S}_{npd}$ is the positive semidefinite matrix *nearest* to $\mathbf{S}$, where nearness is measured with the Frobenius matrix norm. Hence the abbreviation *npd*, nearest positive (semi)definite matrix.

So three steps are needed: (i) compute $\mathbf{S}$ (ii) make it a positive semidefinite matrix using (10) or (11) (iii) compute Glasso using the step two matrix as input. These three steps have been used in Tarr et al. [2015] as well, using the npd algorithm of Higham [2002] and a different choice of $\mathbf{S}$. An advantage of the Gaussian rank correlation (7) is that it is already consistent at the normal distribution, without any additional transformation needed. Then a two-step approach is sufficient.

3 Computation

In this section, we show how easily the sparse and robust precision matrix estimators can be computed in the software package R. In case an estimate of $\mathbf{\Sigma}$ is needed, one simply needs to invert the estimated precision matrix. The function below implements the definitions (10) and (11); the input is a symmetric matrix `sigma` the output a positive semidefinite matrix `sigma.psd`.

```
easy.psd<-function(sigma,method="perturb")
{
if (method=="perturb")
{
p<-ncol(sigma)
eig<-eigen(sigma, symmetric=T, only.values = T)
const<-abs(min(eig$values,0))
sigma.psd<-sigma+diag(p)*const
}
if (method=="npd")
{
eig<-eigen(sigma, symmetric=T)
d<-pmax(eig$values,0)
sigma.psd<-eig$vectors%*%diag(d)%*%t(eig$vectors)
}
return(sigma.psd)
}
```

Assume that the data matrix is in the matrix object `x`. The positive semidefinite matrix **S** based on the transformed Quadrant correlation is computed by the function below

```
quadrant.transformed<-function(x,method="perturb")
{
x.m=apply(x,2,median)
x=sweep(x,2,x.m)
x.s=sign(x)
x.q=apply(x,2,Qn)
cor.quadrant=sin(pi*cor(x.s)/2)
sigma.quadrant=diag(x.q)%*%cor.quadrant%*%diag(x.q)
return(easy.psd(sigma.quadrant,method))
}
```

To compute the Q_n scale estimator, the R-package `robustbase` [Rousseeuw et al., 2015] is needed. For the transformed Spearman correlation we get the corresponding **S** as

```
spearman.transformed<-function(x,method="perturb")
{
x.r=apply(x,2,rank)
x.q=apply(x,2,Qn)
cor.sp=2*sin(pi*cor(x.r)/6)
sigma.sp=diag(x.q)%*%cor.sp%*%diag(x.q)
return(easy.psd(sigma.sp,method))
}
```

The covariance matrix from the Gaussian rank correlations (7) is computed by the function

```
Grank<-function(x)
{
n=nrow(x)
x.q=apply(x,2,Qn)
x.r=apply(x,2,rank)
cor.Grank=cor(qnorm(x.r/(n+1)))
sigma.quadrant=diag(x.q)%*%cor.Grank%*%diag(x.q)
return(sigma.quadrant)
}
```

where we recall that no transformation is needed. The final step is to compute Glasso, with sparsity parameter λ selected by minimizing the BIC criterion (3). The `huge` package of Zhao et al. [2012] allows do this conveniently. The input of the function below is a positive semidefinite matrix `sigma.psd`, and the output a sparse precision matrix estimate.

```
theta.sparse<-function(sigma.psd,n)
{
huge.out<-huge(sigma.psd,method="glasso",verbose=F)
my.bic=-huge.out$loglik+huge.out$df*log(n)/n
opt.i=which.min(my.bic)
return(huge.out$icov[[opt.i]])
}
```

Table 1: Computation time (in sec.) with $n = 50$ averaged over $M = 1000$ samples and over all simulation schemes

	$p = 3$	$p = 30$	$p = 100$
2-step Quadrant	2.02	2.33	4.31
3-step Quadrant (npd)	2.05	2.34	4.35
3-step Quadrant (perturb)	2.04	2.30	4.21
2-step Spearman	2.02	2.32	4.32
3-step Spearman (npd)	2.03	2.32	4.36
3-step Spearman (perturb)	2.00	2.31	4.30
2-step Gaussian Rank	2.02	2.33	4.35
Glasso	1.99	2.33	4.24

For the approach based on the Spearman correlations, for instance, and given a data matrix `x`, the next lines compute the positive semidefinite covariance matrix estimator $\mathbf{S}_{npd}$ and the corresponding precision matrix $\mathbf{\Theta}_{\mathbf{S}_{npd}}$.

```
S.hat=spearman.transformed(x,method="npd")
Theta=theta.sparse(S.hat,n=nrow(x))
```

Table 1 presents computation times for samples of size $n = 50$, averaged over $M = 1000$ simulation runs and over the different sampling distributions used in the Simulation Section 5. Comparing the 2-step and 3-step estimators, one sees that there is only a marginal increase in computation time. Comparing the perturbation method (10) and the nearest positive definite approach (11) one sees that the perturbation method is faster, but the relative difference is marginal. All computation times in Table 1 are relatively close to each other, showing that almost all computation time is taken by computing the Glasso in (2). Note that computation times are increasing with dimension p, but at a rate that seems to be less than linear in p.

4 Breakdown point

A definition of breakdown point appropriate for measuring robustness of high dimensional precision matrices is given in Öllerer and Croux [2014]. Define for any symmetric $p \times p$ matrices $\mathbf{A}$ and $\mathbf{B}$

$$D(\mathbf{A}, \mathbf{B}) = \max\{|\lambda_1(\mathbf{A}) - \lambda_1(\mathbf{B})|, |\lambda_p(\mathbf{A})^{-1} - \lambda_p(\mathbf{B})^{-1}|\},$$

where the ordered eigenvalues of a matrix $\mathbf{A}$ are denoted by $0 \leq \lambda_p(\mathbf{A}) \leq \ldots \leq \lambda_1(\mathbf{A})$. Then the *finite-sample breakdown point under cellwise contamination* of a precision matrix estimate $\hat{\mathbf{\Theta}}$ is defined as

$$\epsilon_n(\hat{\mathbf{\Theta}}, \mathbf{X}) = \min_{m=1,\ldots,n} \{\frac{m}{n} : \sup_{\mathbf{X}^m} D(\hat{\mathbf{\Theta}}(\mathbf{X}), \hat{\mathbf{\Theta}}(\mathbf{X}^m)) = \infty\}, \tag{12}$$

where $\mathbf{X}^m$ denotes a corrupted sample obtained from $\mathbf{X} \in \mathbb{R}^{n\times p}$ by replacing in each column at most m cells by arbitrary values. The following theorem was proven in Öllerer and Croux [2014].

Theorem 1. *The finite sample breakdown point under cellwise contamination of the robust precision matrix estimator* $\hat{\mathbf{\Theta}}_{\mathbf{S}}(\mathbf{X})$ *fulfills*

$$\epsilon_n(\hat{\mathbf{\Theta}}_{\mathbf{S}}, \mathbf{X}) \geq \epsilon_n^+(\mathbf{S}, \mathbf{X}) \tag{13}$$

with $\mathbf{S}$ *a positive semidefinite covariance estimator.*

Here we used the *explosion* finite-sample breakdown point under cellwise contamination of a covariance matrix estimate $\mathbf{S}$, defined as

$$\epsilon_n^+(\mathbf{S}, \mathbf{X}) = \min_{m=1,\ldots,n} \{\frac{m}{n} : \sup_{\mathbf{X}^m} |\lambda_1(\mathbf{S}(\mathbf{X})) - \lambda_1(\mathbf{S}(\mathbf{X}^m))| = \infty\}, \tag{14}$$

where $\mathbf{X}^m$ denotes a corrupted sample obtained from $\mathbf{X}$ by replacing in each column at most m cells by arbitrary values. Theorem 1 shows that Glasso preserves the robustness of the initial estimator. Moreover, Glasso prevents by construction explosion of the precision matrix estimator, and one only needs explosion robustness of the input covariance matrix $\mathbf{S}$.

Consider now our proposal for $\mathbf{S}$, where

$$s_{jk} = \text{scale}(\mathbf{x}^j)\,\text{scale}(\mathbf{x}^k) r(\mathbf{x}^j, \mathbf{x}^k) \qquad j, k = 1, \ldots, p.$$

It was shown in Öllerer and Croux [2014] that the explosion breakdown point under cellwise contamination of $\mathbf{S}$ is always larger than the explosion breakdown point of the scale estimator used. The Q_n-estimator has an explosion breakdown point of 50%, resulting in a breakdown point of 50% under cellwise contamination for the two-step estimators of Section 2. But the correlation measure r in the above definition may be the transformed Quadrant or Spearman correlation given in (8) and (9). In these cases, the three-step estimator discussed in Section 2.2 needs to be used. The following result generalizes Proposition 1 in Öllerer and Croux [2014].

Proposition 1. *Let* $\mathbf{S}$ *be the covariance estimator based on pairwise correlations as defined in (4). Then*

$$\epsilon_n^+(\mathbf{S}_{perturb}, \mathbf{X}) \geq \max_{j=1,\ldots,p} \epsilon_n^+(\text{scale}, \mathbf{x}^j) \quad \textit{and} \quad \epsilon_n^+(\mathbf{S}_{npd}, \mathbf{X}) \geq \max_{j=1,\ldots,p} \epsilon_n^+(\text{scale}, \mathbf{x}^j), \tag{15}$$

with $\epsilon_n^+(\text{scale}, \mathbf{x}^j)$ *the explosion breakdown point of the scale estimator used.*

Proof. We first proof the result for the perturbation method. Using the triangular inequality, we obtain

$$|\lambda_1(\mathbf{S}_{perturb}(\mathbf{X})) - \lambda_1(\mathbf{S}_{perturb}(\mathbf{X}^m))| \leq |\lambda_1(\mathbf{S}_{perturb}(\mathbf{X}))| + |\lambda_1(\mathbf{S}_{perturb}(\mathbf{X}^m))|. \tag{16}$$

From Definition (10) we get

$$\lambda_1(\mathbf{S}_{perturb}(\mathbf{X}^m)) = \lambda_1(\mathbf{S}(\mathbf{X}^m)) - \min(0, \lambda_p(\mathbf{S}(\mathbf{X}^m))). \tag{17}$$

Using a result from Algebra [see Seber, 2008, Equation 6.26a], we have

$$|\lambda_r(\mathbf{S}(\mathbf{X}^m))| \leq p \max_{i,j=1,\ldots,p} |\mathbf{S}(\mathbf{X}^\mathbf{m})_{ij}| \leq p \max_{j,k=1,\ldots,p} \text{scale}((\mathbf{X}^m)^j)\,\text{scale}((\mathbf{X}^m)^k) \tag{18}$$

for all $r = 1, \ldots, p$ and any $m \in \{1, \ldots, n\}$, where $(\mathbf{X}^m)^j$ denotes the jth column of matrix $\mathbf{X}^m$. For the second inequality in (18) we use the fact that the correlation measures (also the transformed ones) have an absolute value smaller than 1.

Equations (16), (17) and (18), together with the definition of the explosion breakdown point (14) show that (15) holds. The proof for the npd method is analogous, and even more simple. Indeed, it follows immediately from (11) that

$$\lambda_1(\mathbf{S}_{npd}(\mathbf{X}^m)) = \lambda_1(\mathbf{S}(\mathbf{X}^m)),$$

where we note that a matrix with non-negative values on the diagonal must have a non-negative largest eigenvalue. □

The proposition above combined with Theorem 1 shows that also the three-stage estimators have an explosion breakdown point under cellwise contamination of at least 50%.

5 Simulations

In this section, we perform a simulation study to compare the performance of the different precision matrix estimators introduced in Section 2. We compare the consistent 3-step estimators of Section 2.2 to the inconsistent 2-step estimator of Section 2.1. For the former, we use both methods for making the symmetric covariance matrix positive semidefinite: the nearest positive definite matrix (npd) method and the perturbation method. We also include the consistent 2-step estimator based on the Gaussian rank correlation (7), for which no third step is needed. As a benchmark, we compare with the nonrobust estimators Glasso, where the sample covariance matrix is taken as an input in (2), and with the inverse of the sample covariance matrix (that can only be computed if $n > p$).

The setup of the simulation study is taken over from Öllerer and Croux [2014]. We use four sampling schemes to cover different patterns of the precision matrix $\mathbf{\Theta}_0 \in \mathbb{R}^{p \times p}$:

- 'banded': $(\mathbf{\Theta}_0)_{ij} = 0.6^{|i-j|}$

- 'sparse': $\mathbf{\Theta}_0 = \mathbf{B} + \delta \mathbf{I}_p$ with $\mathbb{P}[b_{ij} = 0.5] = 0.1$ and $\mathbb{P}[b_{ij} = 0] = 0.9$ for $i \neq j$. The parameter δ is chosen such that the conditional number of $\mathbf{\Theta}_0$ equals p. Then the matrix is standardized to have unit diagonals.
- 'dense': $(\mathbf{\Theta}_0)_{ii} = 1$ and $(\mathbf{\Theta}_0)_{ij} = 0.5$ for $i \neq j$
- 'diagonal': $(\mathbf{\Theta}_0)_{ii} = 1$ and $(\mathbf{\Theta}_0)_{ij} = 0$ for $i \neq j$

For each sampling scheme, we generate $M = 1000$ samples of size $n = 50$ from a multivariate normal $\mathcal{N}(0, \mathbf{\Theta}_0^{-1})$. We take as dimensions $p = 3$, $p = 30$ and $p = 100$. To each data set, we then add 0%, 5% and 10% of cellwise contamination. This means that we randomly select 0%, 5% and 10% of the cells and replace their value with a draw from a normal distribution $\mathcal{N}(10, 0.2)$.

We compare the performance of the different estimators $\hat{\mathbf{\Theta}}$ by the Kullback-Leibler (KL) divergence [see e.g. Bühlmann and van de Geer, 2011]

$$KL(\hat{\mathbf{\Theta}}, \mathbf{\Theta}_0) = \text{tr}(\mathbf{\Theta}_0^{-1}\hat{\mathbf{\Theta}}) - \log\det(\mathbf{\Theta}_0^{-1}\hat{\mathbf{\Theta}}) - p.$$

The lower the value of KL, the better for the estimate. The results for the banded simulation setup are given in Table 2. Let us focus on what is new in this simulation study compared to Öllerer and Croux [2014].

(i) For $p = 3$ and clean data, the inconsistent two-step Quadrant estimator results in a substantially higher KL-value than the consistent three-step Quadrant estimators for clean data and $p = 3$. Here, the additional step leads to a considerable improvement of the estimate. However, in presence of outliers or for higher values of p, the inconsistent two-step Quadrant estimator yields lower values of KL. This is a surprising outcome: rendering the Quadrant correlation based estimator consistent comes at the price of increased Kullback-Leibner distance, at least for the configurations of interest in this paper (n close to or smaller than p).

For the Spearman estimator, there is not much difference between the two-step and three-step method (at least not when using npd). This was to be expected (see Figure 1).

(ii) Comparing the perturbation method and the nearest posive definite (npd) approach, the npd has a clear advantage. Particularly for high dimensions the difference is pronounced (for the Quadrant estimator).

Comparing the different estimators in Table 2 results in the following findings:

(iii) If no outliers are present, and if n is close to p, then Glasso based on the sample covariance matrix is best. But the difference to the 2-step Gaussian rank and the Spearman based estimators is small. The quadrant correlation is much less efficient for clean normally distributed data.

Table 2: Simulation results: Kullback-Leibler criterion (KL) for banded simulation setup with $n = 50$ averaged over $M = 1000$ simulations using BIC criterion to select λ

	$p = 3$			$p = 30$			$p = 100$		
% outliers	0%	5%	10%	0%	5%	10%	0%	5%	10%
2-step Quadrant	0.59	0.83	1.08	11.09	13.35	15.78	38.70	46.91	55.54
3-step Quadrant (npd)	0.33	0.54	0.81	12.43	15.19	18.02	49.54	60.51	71.02
3-step Quadrant (perturb)	0.33	0.54	0.81	16.01	19.44	22.70	63.58	77.43	90.21
2-step Spearman	0.29	0.63	0.98	10.72	13.46	15.92	38.62	47.10	55.66
3-step Spearman (npd)	0.27	0.60	0.96	10.67	13.53	16.06	39.47	48.25	57.01
3-step Spearman (perturb)	0.27	0.60	0.96	10.79	13.56	16.07	40.70	49.80	58.81
2-step Gaussian Rank	0.27	0.68	1.05	10.63	13.50	15.88	38.62	47.12	55.49
Glasso	0.23	2.98	4.12	10.31	30.56	42.48	38.00	106.49	145.56
Sample Covariance	0.14	2.40	3.54	39.45	26.93	31.16			

(iv) Under contamination, the nonrobust Glasso and the sample covariance matrix are not reliable anymore, and have much higher values of KL. For $p = 30$ and $p = 100$, best results are achieved by the two-step Gaussian rank and the two-step estimators based on Quadrant and Spearman correlations. There do not seem to be major differences in performance between the latter three methods in these configurations. This may partly be explained by the fact that the pairwise covariances computed in (4) use the same robust scale estimator.

In the low dimensional setting ($p = 3$) under contamination, Spearman is a bit better than Gaussian Rank, which is on its turn a bit better than Quadrant. For higher levels of contamination (larger than 10 %) this ordering is expected to change in favor of Quadrant correlation.

Result of KL for the other three simulation setups are given in Table 3. For the 'dense' setting exactly the same conclusions can be drawn as for the 'banded' setting of Table 2. For the other two settings, which are characterized by a sparse true precision matrix, we see that Glasso outperforms the sample covariance matrix even for $p = 3$. The overall conclusion of these simulation results is that the two-step Gaussian rank, the two-step Spearman, and the three-step Spearman (npd) are comparable and yield the best results.

6 Graphical models

Sparse estimation of the precision matrix has a direct application in graphical modelling. If element (i, j) of $\hat{\mathbf{\Theta}}$ equals zero, then the estimated partial correlation between variables i and j equals zero. Since we are assuming normality, this means that variables i and j are independent, conditional on the other variables. The variables are represented by the nodes of the graph,

Table 3: Same as Table 2 for the three other simulation setups

		$p = 3$			$p = 30$			$p = 100$		
	% outliers	0%	5%	10%	0%	5%	10%	0%	5%	10%
sparse	2-step Quadrant	0.11	0.25	0.42	7.83	10.19	12.64	36.42	44.85	53.48
	3-step Quadrant (npd)	0.15	0.29	0.45	9.16	12.06	14.81	48.05	59.22	69.60
	3-step Quadrant (perturb)	0.15	0.29	0.45	13.02	16.57	19.78	63.72	77.70	90.41
	2-step Spearman	0.11	0.25	0.42	7.97	10.43	12.86	37.58	45.75	53.97
	3-step Spearman (npd)	0.11	0.26	0.43	7.99	10.53	12.96	38.56	47.07	55.34
	3-step Spearman (perturb)	0.11	0.26	0.43	8.01	10.54	12.97	39.91	48.70	57.38
	2-step Gaussian Rank	0.11	0.25	0.42	8.02	10.42	12.77	37.70	45.55	53.68
	Glasso	0.08	3.15	4.58	7.93	34.43	48.39	37.29	120.93	165.99
	Sample Covariance	0.14	2.70	4.04	39.45	27.16	33.75			
dense	2-step Quadrant	0.56	0.75	0.94	4.39	6.56	8.96	11.36	19.46	28.01
	3-step Quadrant (npd)	0.33	0.52	0.74	5.69	8.32	11.09	21.82	32.77	43.33
	3-step Quadrant (perturb)	0.33	0.52	0.74	8.99	12.33	15.58	35.17	49.70	62.58
	2-step Spearman	0.29	0.67	0.93	4.41	6.58	9.00	11.41	19.34	27.93
	3-step Spearman (npd)	0.25	0.63	0.91	4.48	6.72	9.12	12.15	20.39	29.22
	3-step Spearman (perturb)	0.25	0.63	0.91	4.57	6.72	9.13	13.18	21.83	31.03
	2-step Gaussian Rank	0.25	0.70	0.95	4.40	6.56	8.95	11.39	19.30	27.89
	Glasso	0.20	2.98	4.18	4.22	25.00	37.38	10.73	82.74	123.36
	Sample Covariance	0.14	2.44	3.62	39.45	23.48	26.43			
diagonal	2-step Quadrant	0.11	0.25	0.42	1.92	4.08	6.54	7.83	15.85	24.54
	3-step Quadrant (npd)	0.15	0.29	0.45	3.24	5.85	8.66	18.20	29.34	39.98
	3-step Quadrant (perturb)	0.15	0.29	0.45	6.37	9.82	13.17	31.78	46.00	58.76
	2-step Spearman	0.11	0.25	0.42	1.92	4.08	6.54	7.72	15.82	24.48
	3-step Spearman (npd)	0.11	0.26	0.43	2.01	4.22	6.67	8.50	16.87	25.75
	3-step Spearman (perturb)	0.11	0.26	0.43	2.01	4.22	6.68	9.47	18.29	27.58
	2-step Gaussian Rank	0.11	0.25	0.42	1.92	4.09	6.56	7.70	15.77	24.41
	Glasso	0.08	3.15	4.58	1.75	35.09	50.39	7.04	120.44	173.97
	Sample Covariance	0.14	2.70	4.04	39.45	26.86	34.35			

and if two variables are estimated as conditionally dependent, an undirected arrow is drawn between the corresponding nodes. The rank based correlation coefficient matrices, Spearman or Gaussian rank, can then be used as an input for the Glasso method. Several papers discussed this approach in depth, see Liu et al. [2009], Liu et al. [2012], Xue and Zou [2012] and Zhao et al. [2014b]. They point out an important advantage of using rank-based correlation. If the distribution is only multivariate normal after monotone transformation of the variables (then the distribution is said to be "nonparanormal", and it has a multivariate Gaussian copula), zero partial correlation still implies conditional independence. A major difference with this paper is that we study robust and sparse inverse covariance matrices, and not correlation matrices. Obviously, in the context of graphical modelling, the retrieved graph will be exactly the same.

To measure how well the graph structure is recovered, we compute false positive (FP) and false negative (FN) rates:

$$\text{FP} = \frac{|\{(i,j) : i = 1, \ldots, n; j = 1, \ldots, p : (\hat{\boldsymbol{\Theta}})_{ij} \neq 0 \wedge (\boldsymbol{\Theta}_0)_{ij} = 0\}|}{|\{(i,j) : i = 1, \ldots, n; j = 1, \ldots, p : (\boldsymbol{\Theta}_0)_{ij} = 0\}|}$$

$$\text{FN} = \frac{|\{(i,j) : i = 1, \ldots, n; j = 1, \ldots, p : (\hat{\boldsymbol{\Theta}})_{ij} = 0 \wedge (\boldsymbol{\Theta}_0)_{ij} \neq 0\}|}{|\{(i,j) : i = 1, \ldots, n; j = 1, \ldots, p : (\boldsymbol{\Theta}_0)_{ij} \neq 0\}|}$$

They give the percentage of zero-elements of the precision matrix wrongly estimated as nonzero and the percentage of nonzero-elements that are wrongly estimated as zero. In other words, FN gives the percentage of undetected edges of the graph, and FP the percentage of falsely detected edges. The lower these values are, the better.

To investigate how well the different estimators are able to recover the graph structure, Table 4 gives FP and FN for the setups $p = 30$ and $p = 100$ in the 'sparse' setting. The inverse sample covariance matrix is a nonsparse estimator, and therefore always leads to an FP equal to one and an FN equal to zero. The other estimators lead to pretty similar values of FP and FN. The nonrobust Glasso for $p = 30$ has an increased FN rate under contamination. The 3-step Spearman yields the lowest FP and FN rates in al considered cases, but differences to the other procedures are small.

7 Discussion

We discuss robust and sparse estimators of the precision matrix, computable in high dimensions and for $p > n$. This proceedings paper complements Öllerer and Croux [2014], but we provide further discussion and study additionally the consistent versions of estimators based on Quadrant correlation and Spearman's rank correlation. For computing the latter estimators, an additional step is needed to guarantee positive definiteness of the matrices.

Table 4: Simulation results: False Positive Rate (FP) and False Negative Rate (FN) for the sparse simulation setup with $n = 50$ averaged over $M = 1000$ simulations using BIC criterion to select λ

	$p = 30$						$p = 100$					
% outliers	0%		5%		10%		0%		5%		10%	
	FP	FN	FP	FN	FP	FN	FP	FN	FP	FN	FP	FN
2-step Quadrant	0.01	0.70	0.01	0.71	0.01	0.72	0.00	0.90	0.00	0.90	0.00	0.90
3-step Quadrant (npd)	0.01	0.67	0.01	0.68	0.01	0.70	0.00	0.90	0.00	0.90	0.00	0.90
3-step Quadrant (perturb)	0.00	0.70	0.01	0.71	0.01	0.72	0.00	0.90	0.00	0.90	0.00	0.91
2-step Spearman	0.01	0.64	0.01	0.68	0.01	0.71	0.00	0.89	0.00	0.90	0.00	0.90
3-step Spearman (npd)	0.01	0.63	0.01	0.67	0.01	0.70	0.00	0.89	0.00	0.90	0.00	0.90
3-step Spearman (perturb)	0.01	0.63	0.01	0.67	0.01	0.70	0.00	0.89	0.00	0.90	0.00	0.90
2-step Gaussian Rank	0.01	0.64	0.01	0.69	0.01	0.72	0.00	0.89	0.00	0.90	0.00	0.90
Glasso	0.00	0.64	0.01	0.76	0.01	0.76	0.00	0.89	0.01	0.91	0.00	0.91
Sample Covariance	1.00	0.00	1.00	0.00	1.00	0.00						

We prove that this extra step is not distorting the high breakdown point of the estimators.

The estimators discussed in this paper are using sign and rank correlation measures. Spearman correlation provides a good trade-off between robustness and efficiency. In Croux and Dehon [2010] it was shown that Spearman and Kendall correlation behave rather similarly in the bivariate setting. Using Kendall correlation in the setting of our paper has some disadvantage: (i) there exists a $O(n \log n)$ algorithm to compute it, available in the R-package `pcaPP` [Todorov et al., 2014], but it is still slower than the calculation of a Spearman correlation (ii) The matrix $\mathbf{S}$ one gets from pairwise Kendall correlations is not positive semidefinite, in contrast to Spearman and Quadrant. Even if one is not interested in consistent estimation of $\mathbf{\Sigma}$, a transformation to positive semidefiniteness is required.

While we focused our attention on the estimation of the precision matrix and the covariance matrix, we did not consider the estimation of the location parameter μ yet. Note that estimation based on Spearman correlation does not require an auxiliary location estimate. A simple robust estimator for μ is the coordinatewise median, which simply computes the median for every variable separately. Obviously, this estimator is highly robust and computable in high dimensions. However, this estimator is not affine equivariant, and neither are the covariance matrix estimators $\mathbf{S}$ considered in this paper. If we transform the observation $\mathbf{x_i}$ into $\mathbf{Ax_i} + \mathbf{b}$, with $\mathbf{A}$ a non-singular matrix and $\mathbf{b}$ a constant vector, then the estimators $\hat{\mu}$ and $\hat{\mathbf{\Sigma}}$ are said to be affine equivariant if they change accordingly to $\mathbf{A}\hat{\mu} + \mathbf{b}$ and $\mathbf{A}\hat{\mathbf{\Sigma}}\mathbf{A}^\top$. We only have this property for diagonal matrices $\mathbf{A}$.

A popular robust estimator of location and covariance is the Minimum Covariance Determinant (MCD) estimator [Rousseeuw and Van Driessen, 1999]

which is affine equivariant, but ill defined if $p > n$. Indeed, the MCD is looking for a subsample of half the sample size having smallest value of the determinant of the covariance matrix computed from this subsample. But if $p > n$, or even if $p > n/2$ all the determinants of covariance matrices computed from halfsamples are zero, and it is not clear what to do then. Moreveor, the MCD estimator is not robust to cellwise outliers if you have many of them, as is common in high dimension. There is recent work of Agostinelli et al. [2015] proposing an almost affine equivariant, location/covariance matrix estimator robust to cellwise contamination. Unfortunately, the latter estimator is not computable if $p > n$, as is the proposal of Van Aelst et al. [2011]. To sum up, one needs to give up affine equivariance when constructing robust estimators for $p > n$. We refer to Alqallaf et al. [2009] and Tyler [2010] for further discussion on equivariance properties and contamination models appropriate in high dimensions.

Robust correlation matrices based on pairwise rank correlation estimators have been studied before in the literature. In Section 6 we reviewed their use in graphical modelling. In principal component analysis they have been used by Van Aelst et al. [2010], who used Spearman correlation. Alqallaf et al. [2002] use Quadrant correlation for non-sparse covariance matrix estimation. We believe that the cellwise robust covariance matrix estimators based on ranks and discussed in this and other papers have a lot of potential for high dimensional data analysis.

Acknowledgements

The authors wish to acknowledge the support from the GOA/12/014 project of the Research Fund KU Leuven.

References

C. Agostinelli, A. Leung, V.J. Yohai, and R.H. Zamar. Robust estimation of multivariate location and scatter in the presence of cellwise and casewise contamination. To appear in TEST, 2015. http://arxiv.org/abs/1406.6031.

F.A. Alqallaf, K.P. Konis, R.D. Martin, and R.H. Zamar. Scalable robust covariance and correlation estimates for data mining. In Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 14–23. ACM, 2002.

F.A. Alqallaf, S. Van Aelst, V.J. Yohai, and R.H. Zamar. Propagation of outliers in multivariate data. The Annals of Statistics, 37(1):311–331, 2009.

K. Boudt, J. Cornelissen, and C. Croux. The Gaussian rank correlation

estimator: robustness properties. Statistics and Computing, 22(2):471–483, 2012.

P. Bühlmann and S. van de Geer. Statistics for high-dimensional data. Springer, Heidelberg, 2011.

C. Croux and C. Dehon. Influence functions of the Spearman and Kendall correlation measures. Statistical Methods & Applications, 19(4):497–515, 2010.

J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. Biostatistics, 9(3):432–441, 2008.

R. Gnanadesikan and J.R. Kettenring. Robust estimates, residuals and outlier detection with multiresponse data. Biometrics, 28(1):81–124, 1972.

N.J. Higham. Computing the nearest correlation matrix - a problem from finance. IMA Journal of Numerical Analysis, 22(3):329–343, 2002.

H. Liu, J. Lafferty, and L. Wasserman. The nonparanormal: Semiparametric estimation on high dimensional undirected graphs. Journal of Machine Learning Research, 10:2295–2328, 2009.

H. Liu, F. Han, M. Yuan, J. Lafferty, and L. Wasserman. High-dimensional semiparametric Gaussian copula graphical models. The Annals of Statistics, 40(4):2293–2326, 2012.

R.A. Maronna, R.D. Martin, and V.J. Yohai. Robust Statistics. John Wiley & Sons, Hoboken, New Jersey, 2nd edition, 2006.

V. Öllerer and C. Croux. Robust high-dimensional precision matrix estimation. Technical report, 2014.

P. Rousseeuw, C. Croux, V. Todorov, A. Ruckstuhl, M. Salibian-Barrera, T. Verbeke, M. Koller, and M. Maechler. robustbase: Basic Robust Statistics, 2015. URL `http://CRAN.R-project.org/package=robustbase`. R package version 0.92-3.

P.J. Rousseeuw and C. Croux. Alternatives to the median absolute deviation. Journal of the American Statistical Association, 88(424):1273–1283, 1993.

P.J. Rousseeuw and G. Molenberghs. Transformation of nonpositive semidefinite correlation matrices. Communications in Statistics - Theory and Methods, 22(4):965–984, 1993.

P.J. Rousseeuw and K. Van Driessen. A fast algorithm for the minimum covariance determinant estimator. Technometrics, 41(3):212–223, 1999.

G.A.F. Seber. A matrix handbook for Statisticians. John Wiley & Sons, Hoboken, New Jersey, 2008.

G. Tarr, S. Müller, and N.C. Weber. Robust estimation of precision matrices under cellwise contamination. Computational Statistics & Data Analysis, 2015. doi: http://dx.doi.org/10.1016/j.csda.2015.02.005.

V. Todorov, P. Filzmoser, H. Fritz, and K. Kalcher. pcaPP: Robust PCA by Projection Pursuit, 2014. URL `http://CRAN.R-project.org/package=pcaPP`. R package version 1.9-60.

D. Tyler. A note on multivariate location and scatter statistics for sparse data. Statistics & Probability Letters, 80(17-18):1409–1413, 2010.

S. Van Aelst, E. Vandervieren, and G. Willems. Robust principal component analysis based on pairwise correlation estimators. In Proceedings of COMPSTAT2010, pages 573–580. Physica-Verlag HD, 2010.

S. Van Aelst, E. Vandervieren, and G. Willems. Stahel-Donoho estimators with cellwise weights. Journal of Statistical Computation and Simulation, 81(1):1–27, 2011.

L. Xue and H. Zou. Regularized rank-based estimation of high-dimensional nonparanormal graphical models. The Annals of Statistics, 40(5):2541–2571, 2012.

M. Yuan and Y. Lin. Model selection and estimation in the Gaussian graphical model. Biometrika, 94(1):19–35, 2007.

T. Zhao, H. Liu, K. Roeder, J. Lafferty, and L. Wasserman. The `huge` package for high-dimensional undirected graph estimation in R. Journal of Machine Learning Research, 13:1059–1062, 2012.

T. Zhao, H. Liu, K. Roeder, J. Lafferty, and L. Wasserman. huge: High-dimensional undirected graph estimation, 2014a. URL `http://CRAN.R-project.org/package=huge`. R package version 1.2.6.

T. Zhao, K. Roeder, and H. Liu. Positive semidefinite rank-based correlation matrix estimation with application to semiparametric graph estimation. Journal of Computational and Graphical Statistics, 23(4):895–922, 2014b.

FACULTY OF ECONOMICS AND BUSINESS
Naamsestraat 69 bus 3500
3000 LEUVEN, BELGIË
tel. + 32 16 32 66 12
fax + 32 16 32 67 91
info@econ.kuleuven.be
www.econ.kuleuven.be